ĐẠI HỌC BÁCH KHOA

25 YEARS ANNIVERSARY

SOICT

ĐẠI HỌC BÁCH KHOA HÀ NỘI
VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

Nhập môn
Khoa học dữ liệu
(IT4930)

Cấu trúc môn học

- Số tuần: 17
  - Lý thuyết: 11-12 tuần
  - Sinh viên trình bày tiến độ công việc đồ án môn học: 01 tuần
  - Sinh viên trình bày đồ án môn học: 03-04 tuần
- Thời gian và địa điểm

- FB Group: https://www.facebook.com/groups/230365807881078/
- Thời gian gặp sinh viên
  - **Hẹn trước qua e-mail**
  - Nhà B1

Môn học này

- Bạn sẽ học cách lấy dữ liệu để
  - Hiểu
  - Xử lý
  - Trích xuất giá trị
  - Trực quan hoá
  - Chia sẻ với người khác
  - Tạo các phán đoán

"The ability to take **data** – to be able to **understand** it, to **process** it, to **extract value** from it, to **visualize** it, to **communicate** it's going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids. Because now we really do have essentially free and ubiquitous data."
- Hal Varian, Google's Chief Economist

Nội dung môn học

- Lecture 1: Tổng quan về Khoa học dữ liệu
- Lecture 2: Thu thập và tiền xử lý dữ liệu
- Lecture 3: Làm sạch và tích hợp dữ liệu
- Lecture 4: Phân tích và khám phá dữ liệu
- Lecture 5: Trực quan hoá dữ liệu
- Lecture 6: Trực quan hoá dữ liệu đa biến
- Lecture 7: Học máy
- Lecture 8: Phân tích dữ liệu lớn
- Lecture 9: Báo cáo tiến độ bài tập lớn và hướng dẫn
- Lecture 10+11: Phân tích một số kiểu dữ liệu
- Lecture 12: Đánh giá kết quả phân tích

Các thầy cô giảng dạy

Khoat Than

Viet-Trung Tran

Huyen-Trang Pham (TA)

Oanh Nguyen

Mai-Anh Bui

Thư viện hoặc ngôn ngữ

python™

ANACONDA
Powered by Continuum Analytics®

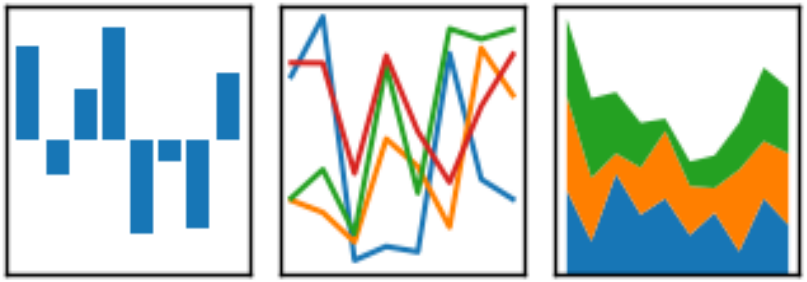
APACHE Spark™

TensorFlow

PyTorch

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



Đánh giá

- Đồ án môn học (**P**): Tối đa 10 điểm
  - Mỗi đồ án được thực hiện bởi một nhóm sinh viên
  - Chọn bài toán thực tế muốn giải quyết
  - Chọn một phương pháp phân tích dữ liệu để giải quyết một bài toán thực tế
  - Cài đặt và đánh giá hiệu năng của phương pháp đó dựa trên dữ liệu thực tế
- Thi trắc nghiệm (**E**): Tối đa 10 điểm
- Điểm học phần (**G**)
  - **G** = 0,4 x **P** + 0,6 x **E**

Đồ án môn học: đề tài

- Tự do đề xuất bài toán thực tế, (các) giải thuật để giải quyết bài toán, và (các) tập dữ liệu được sử dụng
- Đề xuất đề tài phải được **diễn giải cụ thể**
  - **Mô tả bài toán thực tế** sẽ được giải quyết (mục đích, yêu cầu, kịch bản ứng dụng, …)
  - Xác định rõ **giải thuật** dự kiến dùng để giải quyết bài toán.
  - Trình bày các thông tin về **đầu vào (input)** và **đầu ra (output)** của hệ thống học máy sẽ được cài đặt, và **cách thức biểu diễn dữ liệu.**
  - Xác định rõ **(các) tập dữ liệu (datasets)** sẽ được sử dụng.

Đồ án môn học: các yêu cầu

- Kết quả của đồ án phải được trình bày ở cuối môn học

  Tất cả các thành viên phải tham gia vào việc thực hiện và trình bày đồ án

- Báo cáo kết quả của đồ án bao gồm:
  - **Mã nguồn** (source codes): lưu trong một file nén
  - **File hướng dẫn** (readme.txt) mô tả chi tiết cách thức cài đặt/biên dịch/chạy chương trình (và các gói phần mềm được sử dụng kèm theo)
  - **Tài liệu báo cáo** kết quả đồ án mô học (lưu trong file .pdf):
    - Giới thiệu và mô tả về bài toán thực tế được giải quyết
    - Các chi tiết của (các) phương pháp phân tích và (các) tập dữ liệu được sử dụng
    - Các kết quả thí nghiệm đánh giá hiệu quả hoặc kết quả phân tích
    - Các chức năng chính của hệ thống (và cách sử dụng)
    - Cấu trúc của mã nguồn chương trình, vai trò của các lớp (classes) và các phương thức (methods) chính/quan trọng
    - Các vấn đề/khó khăn gặp phải trong quá trình thực hiện công việc của đồ án, và cách thức được dùng để giải quyết (vượt qua)
    - Các khám phá mới hoặc kết luận

Đồ án môn học: đánh giá

- Công việc đồ án được đánh giá theo các tiêu chí sau:
  - *Mức độ phức tạp / khó khăn của bài toán thực tế được giải quyết*
  - *Chất lượng (sự đúng đắn và phù hợp) của phương pháp được dùng để giải quyết bài toán*
  - *Đánh giá và lựa chọn kỹ lưỡng mô hình*
  - Chất lượng của bài trình bày (presentation) kết quả đồ án
  - Chất lượng của tài liệu báo cáo kết quả đồ án
  - Cài đặt hệ thống thử nghiệm (các chức năng, dễ sử dụng, …)
- Bài trình bày trong khoảng 15 phút, và phù hợp với những gì được nêu trong tài liệu báo cáo
- **Nếu sử dụng lại / kế thừa / khai thác các mã nguồn / các gói phần mềm / các công cụ sẵn có, thì phải nêu rõ ràng và chính xác trong tài liệu báo cáo (và đề cập trong bài trình bày)**

Tài liệu học tập

- Reference books:
  - Grus, Joel. *Data science from scratch: first principles with python*. O'Reilly Media, Inc., 2015.
  - Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2017.
  - Han, Jiawei, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

ĐẠI HỌC BÁCH KHOA

25 YEARS ANNIVERSARY SOICT

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

Thank you for your attentions!

soict.hust.edu.vn/ fb.com/groups/soict